entropy

Article

Machine Learning Predictors of Extreme Events Occurring in Complex Dynamical Systems

Stephen Guth and Themistoklis P. Sapsis *

Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave.,
Cambridge, MA 02139, USA; sguth@mit.edu
* Correspondence: sapsis@mit.edu; Tel.:+617-324-7508; Fax: +617-253-8689

Received: 22 August 2019; Accepted: 20 September 2019; Published: 23 September 2019

Abstract: The ability to characterize and predict extreme events is a vital topic in fields ranging from finance to ocean engineering. Typically, the most-extreme events are also the most-rare, and it is this property that makes data collection and direct simulation challenging. We consider the problem of deriving optimal predictors of extremes directly from data characterizing a complex system, by formulating the problem in the context of binary classification. Specifically, we assume that a training dataset consists of: (i) indicator time series specifying on whether or not an extreme event occurs; and (ii) observables time series, which are employed to formulate efficient predictors. We employ and assess standard binary classification criteria for the selection of optimal predictors, such as total and balanced error and area under the curve, in the context of extreme event prediction. For physical systems for which there is sufficient separation between the extreme and regular events, i.e., extremes are distinguishably larger compared with regular events, we prove the existence of optimal extreme event thresholds that lead to efficient predictors. Moreover, motivated by the special character of extreme events, i.e., the very low rate of occurrence, we formulate a new objective function for the selection of predictors. This objective is constructed from the same principles as receiver operating characteristic curves, and exhibits a geometric connection to the regime separation property. We demonstrate the application of the new selection criterion to the advance prediction of intermittent extreme events in two challenging complex systems: the Majda–McLaughlin–Tabak model, a 1D nonlinear, dispersive wave model, and the 2D Kolmogorov flow model, which exhibits extreme dissipation events.

Keywords: optimal predictors; binary classification; rare extreme events; chaotic systems; data-driven methods

1. Introduction

Many phenomena in a wide range of physical domains and engineering applications have observable properties that are normally distributed, that is, they obey Gaussian statistics. Gaussian-distributed random variables and processes are particularly easy to manipulate algebraically, and there is a rich literature using their properties in widely varying areas of probability and statistics, from Bayesian regression [1] to stochastic differential equations [2]. In many applications, however, random variables have significantly non-Gaussian character. Frequently, the "long-tails" of their distribution, which contain extreme, but rare events, are particularly important for a complete understanding of the phenomena in question. Examples of this behavior occur in systems ranging from ocean waves [3,4] to finance [5], but similar phenomena are observed in fields as far afoot as cell dynamics [6], mechanical part failure [7], and turbine shocks [8].

Several approaches have been developed that successfully resolve the statistics capturing extreme events (see, e.g., [9–14]). These often take advantage of the special structure of the problem, which is of course the preferred approach in cases where the dynamics are well understood. For systems where our understanding of the governing laws is partial or in situations where there are important model errors, it can be shown that combining the available imperfect models with data-driven ideas can result in prediction schemes that are more effective than each of the ingredients, data or imperfect model, used independently [15]. Another perspective, for systems where we can control which samples to use or compute, is to apply optimal experimental design ideas or active learning [16]. In this case, one can optimize the samples that should be used, using information from samples already available, in order to have accurate statistics with a very small computational or experimental cost.

There are situations, however, where the only information available that characterizes the system is observations—"big data". For such cases, one important class of problems is to identify extreme event precursors—system states that are likely to evolve into extreme events. Successful identification of precursors requires both a careful definition of what exactly qualifies as an extreme event, as well as a balance between false positive and false negative errors [17–22]. While there is already a vast literature in machine learning related to binary classification, little is specifically directed to the problem of extreme event precursors and prediction.

In this paper, we first discuss limitations of standard methods from binary classification, in the context of extreme event precursors. Motivated by these limitations, we design a machine learning approach to compute optimal precursors to extreme events ("predictors") directly from data, taking into account the most important aspect of extreme events: their rare character. This approach will naturally suggest a frontier for trade-offs between false positive and negative error rates, and in many cases will geometrically identify an optimal threshold to separate extreme and quiescent events. We demonstrate the derived ideas to two prototype systems that display highly complex chaotic behavior with elements of extreme, rare events: the Majda–McLaughlin–Tabak model [23] and the Kolmogorov flow [24]. In the first system, we compare our method to a standard loss function (total error rate) and show how it leads to better predictors. In both systems, machine-learned predictors support previous analysis on the mechanisms of intermittency in those systems.

2. A Critical Overview of Binary Classification Methods

The problem of identifying good precursors for extreme events can also be seen as a binary classification problem. In this case, a training dataset takes the form of a set of pairs $S = \{(a_i, \mathbf{x}_i)\}$, where a_i is a scalar indicator or quantity of interest at time t_i, whose value defines whether or not we have an extreme event. The vector $\mathbf{x}_i$, on the other hand, contains all the possible observables available up to time $t_i - \tau$ that we can utilize to predict whether we have an extreme event or not in the near future, i.e., after time τ. The aim is to identify the most effective function $b = b(\mathbf{x})$ so that using the value of b we can predict if a will exhibit extreme behavior.

For any given function $b(\mathbf{x})$, the set of these predictor–indicators pairs together defines a two dimensional joint distribution, with probability density function (pdf) $f_{ab}(a, b)$ and cumulative distribution function (cdf)

$$F_{ab}(\hat{a}, \hat{b}) = P(a < \hat{a}, b < \hat{b}). \tag{1}$$

We use hat notation when the choice of $\hat{a}$ or $\hat{b}$ corresponds to a threshold value. A value exceeding the threshold is called extreme, and a value not exceeding it is called quiescent. The joint distribution may be constructed from simulations, from experimental measurements, or from analytical models. We use the term *histogram* in this work to refer to the data and their functional representation as a probability distribution.

A fixed choice of $\hat{a}$ and $\hat{b}$ defines a binary classification problem with four possibilities (Figure 1):

- True Positive (TP): an event predicted to be extreme $(b > \hat{b})$ that is actually extreme $(a > \hat{a})$;
- True Negative (TN): an event predicted to be quiescent $(b < \hat{b})$ that is actually quiescent $(a < \hat{a})$;
- False Positive (FP): an event predicted to be extreme $(b > \hat{b})$ that is actually quiescent $(a < \hat{a})$;
- False Negative (FN): an event predicted to be quiescent $(b < \hat{b})$ that is actually extreme $(a > \hat{a})$.

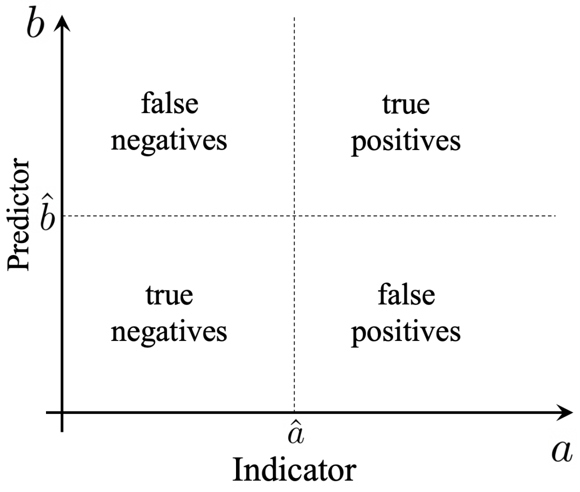


Figure 1. Optimization of predictors as a binary classification problem.

2.1. Total and Balanced Error Rate

Based on this classification, we can define several criteria for the selection of predictors. The typical binary classification task minimizes the *total error rate*, which is defined as

$$E_T = P(FP) + P(FN) = P(a < \hat{a}, b > \hat{b}) + P(a > \hat{a}, b < \hat{b}) \tag{2}$$

This error metric is poorly suited for the extreme event prediction problem for two reasons:

- First, total error rate is unsuited for unbalanced data. Extreme events are usually associated with extremely unbalanced datasets. This manifests in two ways. First, even a naive predictor may achieve very high accuracy, simply because always predicting "not extreme" is usually correct. Second, resampling the data (for instance, to balance the number of extreme and not-extreme training points) may widely change the total error rate, which in turn may change the optimal predictor.
- Second, this error metric is unsuited for strength-of-confidence measurement. It has no ability to distinguish between confidently classified points and and un-confidently classified points. This is particularly important if we expect our predictor to make many mistakes.

The first objection may be resolved by using balanced error quantities, such as the balanced error rate:

$$E_B = 1 - \frac{1}{2}\left(\frac{P(TP)}{P(TP)+P(FN)} + \frac{P(TN)}{P(TN)+P(FP)}\right), \tag{3}$$

which measures the true-positives and true-negatives but normalized by the number of predicted positives and negatives, respectively.

2.2. F_1-Score

Another criterion to deal with the strongly unbalanced character of the datasets that contain extreme rare events is the F_1-score. We first define several other important quantities that we use and study their basic properties.

In particular, precision, denoted by s, is the probability that an event is a true positive, given that it is predicted to be extreme:

$$s(\hat{a},\hat{b}) = \frac{P(TP)}{P(TP)+P(FP)} = \frac{1+F_{ab}(\hat{a},\hat{b}) - F_a(\hat{a}) - F_b(\hat{b})}{1-F_b(\hat{b})} \quad \text{[Precision]}. \tag{4}$$

The recall (sometimes called sensitivity), denoted by r, is the probability that an event is a true positive, given that is actually extreme:

$$r(\hat{a}, \hat{b}) = \frac{P(TP)}{P(TP) + P(FN)} = \frac{1 + F_{ab}(\hat{a}, \hat{b}) - F_a(\hat{a}) - F_b(\hat{b})}{1 - F_a(\hat{a})} \text{ [Recall]}, \tag{5}$$

Furthermore, the extreme event rate, denoted by q, is the probability that an event is actually extreme:

$$q(\hat{a}) = P(TP) + P(FN) = P(a > \hat{a}) = 1 - F_a(\hat{a}) \text{ [Extreme event rate]}.$$

By construction, the derived quantities have the following monotonicity properties:

Theorem 1 (Monotonicity)**.** *The precision function, $s(\hat{a}, \hat{b})$, is monotonic in its first argument, $\hat{a}$, while the recall function, $r(\hat{a}, \hat{b})$, is monotonic in its second argument, $\hat{b}$. Furthermore, the extreme event rate, $q(\hat{a})$, is a monotonic function with respect to its argument, $\hat{a}$.*

Additionally, the choice of precision and recall as binary classification metrics is motivated by the following three invariance properties

Theorem 2 (Invariance-I)**.** *Precision, recall, and extreme event rate are entirely determined by the function F_{ab}, and the thresholds $\hat{a}$ and $\hat{b}$.*

Theorem 3 (Invariance-II)**.** *Let $f_1(a, b)$ and $f_2(a, b)$ be two histograms with the property that, outside of a certain region $[-\infty, a^*] \otimes [-\infty, b^*]$, $f_1(a, b) = \beta f_2(a, b)$ for some fixed constant β. That is to say, $f_1(a, b)$ and $f_2(a, b)$ correspond to histograms that differ only by some number of points (a_i, b_i) for which $a_i < a^*$ and $b_i < b^*$.*

Then, for all $\hat{a} > a^*$ and $\hat{b} > b^*$, $s(\hat{a}, \hat{b}; F_1) = s(\hat{a}, \hat{b}; F_2)$ and $r(\hat{a}, \hat{b}; F_1) = r(\hat{a}, \hat{b}; F_2)$.

Theorem 4 (Invariance-III)**.** *Let $h_1(y), h_2(y) : \mathcal{R} \to \mathcal{R}$ be order-preserving monotonic functions. Let F_{ab} be a histogram, and let $F_{a'b'}$ be the histogram formed from the dataset $\{(h_1(a_i), h_2(b_i))\}$. Then,*

$$\begin{aligned} s(\hat{a}, \hat{b}; F_{ab}) &= s(h_1(\hat{a}), h_2(\hat{b}); F_{a'b'}) \\ r(\hat{a}, \hat{b}; F_{ab}) &= r(h_1(\hat{a}), h_2(\hat{b}); F_{a'b'}) \\ q(\hat{a}; F_{ab}) &= q(h_1(\hat{a}); F_{a'b'}) \end{aligned} \tag{6}$$

In other words, precision, recall, and extreme event rate are invariant under arbitrary nonlinear monotonic rescalings of the indicator and predictor.

Proof. The proof of Theorem 2 follows directly from the definitions of q, r, and s in Equation (6), where they are expressed in terms of conditional probabilities.

The proof of Theorem 3 follows from the restatement of the definition in Equation (6) in the form of ratios of cumulative density functions. The numerators can be written in cumulative distribution form as $1 + F_{ab}(\hat{a}, \hat{b}) - F_a(\hat{a}) - F_b(\hat{b}) = \int_{\hat{b}}^{\infty} \int_{\hat{a}}^{\infty} f_{ab}(a, b) da db$, which takes as support f_{ab} only in the complement of the region $[-\infty, a^*] \otimes [-\infty, b^*]$. Similarly, the denominator support does not intersect the variable region. Because s and r are ratios of definite integrals, each of which is linear in its integrand, they are invariant under the linear rescaling β. Note that this is not true for q, which is not defined as a ratio.

Finally, the proof of Theorem 4 follows from the familiar u-substitution rule of ordinary calculus, restricted to the well behaved class of monotonic substitutions. □

These invariance properties simplify the issue of scale in the choice of predictor functions, limiting the hypothesis space of potential predictors. Using the precision and recall function, we define the the $F-$*score*, given by the harmonic mean of the precision and the recall:

$$F_1 = \frac{2}{r^{-1} + s^{-1}}. \tag{7}$$

Both the F_1-score and the balanced error depend on normalized quantities which take into account the extremely unbalanced character of the datasets. However, their value depends on the thresholds $\hat{a}$ and $\hat{b}$.

2.3. Area under the Precision–Recall Curve

To overcome the dependence on the $\hat{b}-$threshold value, we we consider a fixed extreme event rate, $q(\hat{a})$, or equivalently fixed $\hat{a}$, and define the *precision–recall curve* (*sr* curve) as

$$\rho(\hat{b}; \hat{a}) = \big(r(\hat{a}, \hat{b}), s(\hat{a}, \hat{b})\big) \tag{8}$$

Because $r(\hat{a}, \hat{b})$ is invertible in its second argument (Theorem 1), this curve gives precision as a unique function of recall and the extreme event rate:

$$s = s(r; q) \tag{9}$$

An example *sr* curve (for fixed q) is exhibited in Figure 2. Smaller values of recall correspond to larger values of precision and vice versa. This is intuitive: to be sure to catch every extreme event (high recall), the predictor will have to let through many false positive quiescent events (low precision). Note the precision does not vanish as r increases, even when the predictor threshold, $\hat{b}$, is arbitrarily small. Instead, if that is the case, all events will be predicted as extremes and therefore the precision of the predictor will be given by the rate of extreme events. Therefore, the following limit holds:

Theorem 5 (Extreme Event Rate Correspondence). *Let the sr curve, ρ, correspond to extreme event rate q. Then, we have the following limit*

$$\lim_{r \to 1} s(r; q) = q \tag{10}$$

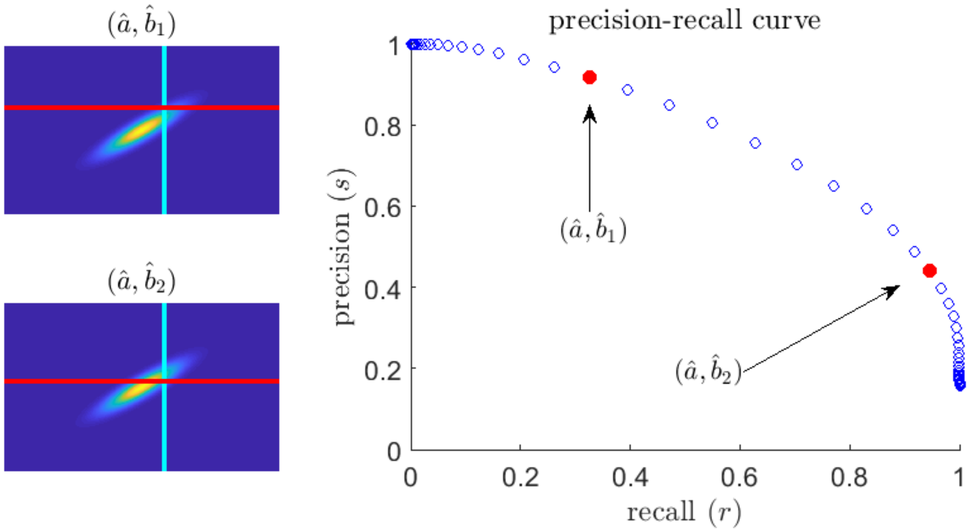


Figure 2. (**left**) Plot of a sample pdf f_{ab}; (**right**) the corresponding precision–recall *sr* curve. The parameterized curve is generated by fixing $\hat{a}$ and letting $\hat{b}$ vary.

The sr curve is a variation of the Receiver Operating Characteristic (ROC) curve used in medical literature [19], which displays predictor performance over a range of threshold values. Here, we employ the sr curve as various reviews (see, e.g., [18]) have shown that it performs better on discriminating between classifiers when the data are wildly unbalanced. To obtain a metric that does not depend on a specific threshold value of the predictor, $\hat{b}$, a standard metric is the area under the curve (AUC), α.

In particular, for a fixed value of the rate q (equivalent to a fixed extreme event threshold $\hat{a}$), the area under the curve, α, is given by

$$\alpha(q) = \int_0^1 s(r)dr = \int_{-\infty}^{\infty} s(\hat{b}) \left| \frac{\partial r}{\partial \hat{b}} \right| d\hat{b}. \tag{11}$$

A larger value of α corresponds to a more favorable set of choices $\hat{b}$ that maximize both precision and recall in combination. The ideal rs curve would run from $(0,1)$ to $(1,1)$, and then down from $(1,1)$ to $(1,q)$. Therefore, we can choose a predictor by maximizing α for a fixed value of q. The value of q, however, has to be chosen in an ad-hoc manner and this motivates the introduction of the next measure.

2.4. Volume under the Precision–Recall–Rate Surface

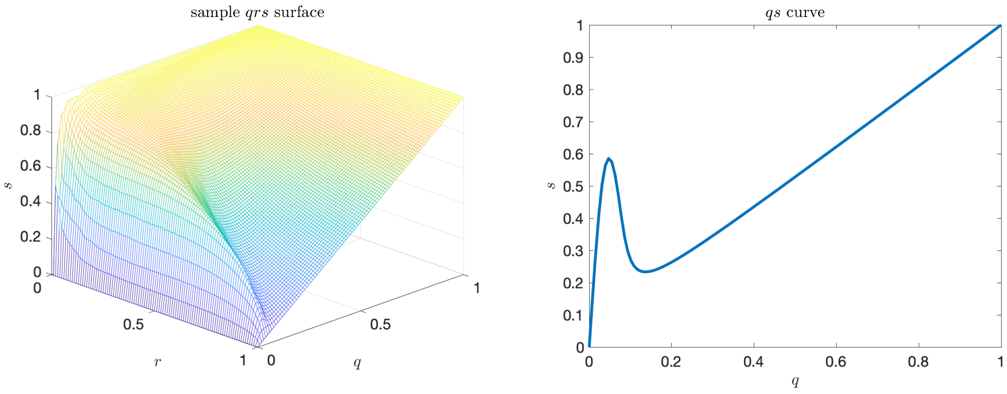
To overcome the dependence on the ad-hoc parameter, $\hat{a}$, we generalize the notion of the precision–recall curve to the precision–recall–rate surface (qrs surface). This is the parametric surface defined by

$$\sigma(\hat{a},\hat{b}) = \big(q(\hat{a}), r(\hat{a},\hat{b}), s(\hat{a},\hat{b})\big). \tag{12}$$

Such qrs surface is shown in Figure 3a. Similar to the sr curve, q and r may be inverted sequentially. By analogy with the area under the curve, α, we can define an enclosed volume functional for the qrs surface as well. In particular, we have the *volume under the surface*, V, given by

$$V = \int_0^1 \int_0^1 s(r,q)drdq. \tag{13}$$

While V evaluates the goodness of a predictor over all possible pairs $(\hat{a},\hat{b})$, it does not specifically quantify the quality of the predictor for extreme-events, i.e., low values of the rate, q.



(**a**) (**b**)

Figure 3. (**a**) Plot of a sample precision–recall–extreme-event-rate, qrs surface. The surface is generated by varying $\hat{a}$ and $\hat{b}$; (**b**) Precision–rate slice of the qrs plot, where $r = 0.5$.

3. Separation of Extreme and Quiescent Events

Before we proceed to the definition of a measure that explicitly takes into account the rare event character (or low rate) of extreme events, it is important to study some properties of the precision–recall–rate surface. We are interested in understanding the structure of these surfaces and the implications on selecting predictors, for physical systems exhibiting extreme events. For a large class of such systems, the causal mechanism of extremes are internal instabilities, combined with the intrinsic stochasticity of the system (see, e.g., [14,20,25–27]). In this case, the extreme events are distinguishably larger compared with regular events. This is because extreme event properties are primarily controlled by the system nonlinearity and the subsequent instabilities. For such cases, the joint pdf between indicator and predictor has a special structure where two distinguished probability regions can be observed. A sample pdf exhibiting this property is shown in Figure 4.

We have observed that this separation of regimes in the probabilistic sense implies interesting properties for the qrs surface (shown in Figure 3 for the pdf shown in Figure 4). In particular, the separation in the pdf f_{ab} results in a knuckle point on the qs curves (Figure 3b). This knuckle point is associated with a local maximum of the precision for the particular choice of the rate, q. It essentially defines an optimal value of the extreme event threshold, $\hat{a}$, which, if chosen, optimizes the performance of the predictor.

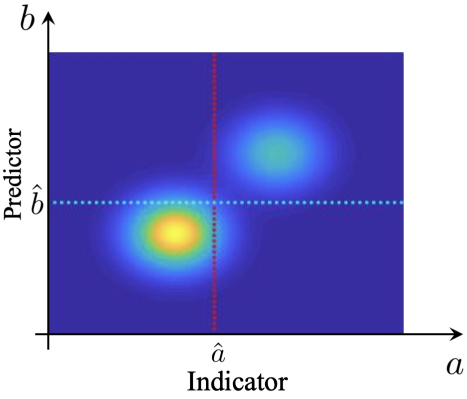


Figure 4. Sample pdf exhibiting separation of extremes from quiescent events, corresponding to the qrs surface in Figure 3a.

Motivated by these observations, we study further the topological properties of the qrs surface. We have already presented several properties of the rs curves and to this end we focus on the qs curves. We first note the following basic properties:

Theorem 6 (Basic qs curve properties)**.** *Let f_{ab} be a continuous joint pdf for the indicator and predictor, with finite support. Then, for all values of recall, r, the qs curve has the following properties:*

1. $0 \leq s(q,r) \leq 1$, *for all* q, r;
2. $s(q=0) = 0$ *and* $s(q=1) = 1$; *and*
3. $\frac{\partial s}{\partial q}(q=0) \geq 0$ *and* $\frac{\partial s}{\partial q}(q=1) \geq 0$.

Proof. Property 1 follows from the cdf form of Definition 4, which is expressed as a ratio. Both numerator and denominator are necessarily positive, leading to $0 \leq s(q,r)$. Furthermore, the numerator is written as the denominator minus a positive function, thus the numerator is never larger than the denominator, leading to $s(q,r) \leq 1$.

Property 2 follows from the confusion matrix formulation of Definition 4. When $q = 0$, there are no true positive events, thus the numerator of s is identically zero. When $q = 1$, there are no false positive events, so the numerator and denominator are equal.

Property 3 follows from the requirements of previous two properties and continuity. □

The above properties restrict the possible shapes of the qs curves. We are interested to understand conditions that lead to the occurrence of a knuckle point. Clearly, this is going to be the case when we have an interval of rates for which $\frac{\partial s}{\partial q} < 0$. We compute this condition in terms of the joint pdf f_{ab}. We have the derivative of s with respect to q for constant r:

$$\left.\frac{ds}{dq}\right|_r = \frac{1}{\frac{\partial q}{\partial \hat{a}}\frac{\partial r}{\partial \hat{b}}}\left(\frac{\partial s}{\partial \hat{a}}\frac{\partial r}{\partial \hat{b}} - \frac{\partial s}{\partial \hat{b}}\frac{\partial r}{\partial \hat{a}}\right). \tag{14}$$

Substituting the expressions for the precision and recall function, we find that the sign of the above derivative is given by the function

$$\begin{aligned} Q(\hat{a},\hat{b}) = &\int_{\hat{a}}^{\infty}\int_{-\infty}^{\infty} f_{ab}(a,b)dbda \int_{\hat{b}}^{\infty} f_{ab}(\hat{a},b)db \int_{-\infty}^{\infty} f_{ab}(a,\hat{b})da \\ &+ \int_{-\infty}^{\infty}\int_{\hat{b}}^{\infty} f_{ab}(a,b)dbda \int_{-\infty}^{\infty} f_{ab}(\hat{a},b)db \int_{\hat{a}}^{\infty} f_{ab}(a,\hat{b})da \\ &- \int_{\hat{a}}^{\infty}\int_{\hat{b}}^{\infty} f_{ab}(a,b)dbda \int_{-\infty}^{\infty} f_{ab}(\hat{a},b)db \int_{-\infty}^{\infty} f_{ab}(a,\hat{b})da \end{aligned} \tag{15}$$

The first two terms of Q will always be positive. That implies that Q can only change sign when the third term grows large enough in magnitude to dominate the other two terms. It is straightforward to show that this will be the case if there are threshold values such that the last term is larger than both the first and second term, i.e., the following sufficient condition holds for some threshold values:

$$s(\hat{a},\hat{b}) > P(b > \hat{b}|a = \hat{a}) \quad and \quad r(\hat{a},\hat{b}) > P(a > \hat{a}|b = \hat{b}) \tag{16}$$

The first condition requires that the ratio of the true positives, $P(TP)$, over all extremes, $P(TP) + P(FP)$, is larger than the the conditional probability for a positive prediction $(b > \hat{b})$ given that the indicator is just going to reach the threshold value, $\hat{a}$. The second condition requires that the ratio of the true positives, $P(TP)$, over all extreme predictions, $P(TP) + P(FN)$, is larger than the the conditional probability for an extreme $(a > \hat{a})$ given that the predictor is just reaching the threshold value, $\hat{b}$. If both of these conditions hold, then there is an optimal value of $\hat{a}$ or q for which the precision of the predictor is optimized.

4. A Predictor Selection Criterion Adjusted for Extreme Events

We introduce a new criterion for optimizing predictors, which takes into account the special features of extreme events, in particular their low rate of occurrence. This criterion should also be able to distinguish predictors that optimally capture extreme events for the case where we have well separated extreme events (knuckle point in the qs curve) but also in more typical cases.

We begin by noting that a completely uninformed predictor (e.g., flip a weighted coin to determine the prediction) will have the parametric form $s = q$. This is because such predictor will only need to be consistent with the rate of occurrence of extreme events. Any other predictor will have $s > q$ at least somewhere. Note that if we find a predictor with $s < q$ this can be transformed into a good one by just taking the opposite guess. To this end, it is meaningful to consider the difference between the AUC of any arbitrary predictor, $\alpha(q)$, and the completely uninformed one which has AUC equal to q. That is, we should take the amount by which the predictor is better than an appropriately weighted coin, in the sense of area under the precision–recall curve. Moreover, $\alpha(q)$ will vary in value as q is varied. Our extreme event predictor should single out one specific threshold that corresponds to the most effective separation (between extremes and quiescent events), the threshold that corresponds to the best predictive power.

Based on this analysis, we suggest the maximum adjusted area under the curve, α^*, as a choice of metric, given by

$$\alpha^* = \max_{q\in[0,1]} (\alpha(q) - q). \tag{17}$$

The quantity $\alpha(q) - q$ is a measure, at extreme event rate q, of how much better a predictor B is than the uninformed predictor. When $\alpha(q) - q \gg 0$, the predictor does an excellent job of predicting extreme events at the threshold $\hat{a}$ corresponding to the extreme event rate q. Conversely, when $\alpha(q) - q \approx 0$ (or even $\alpha(q) - q < 0$), the predictor is poor at that extreme event rate. Below, we examine several analytical forms of joint pdfs between indicators and predictors, to understand the mechanics of the volume under the surface metric and the maximum adjusted area under the curve.

4.1. Coinflip Indicator—Predictor

The coinflip predictor is the naive predictor that is completely independent of the indicator. Its characteristic triangular shape, shown in Figure 5, has the minimum value of the volume under the surface: $V = 0.5$. Further, the coinflip predictor is the baseline against which the metric α^* is constructed. Indeed, for the coinflip predictor, $\alpha^* = 0$.

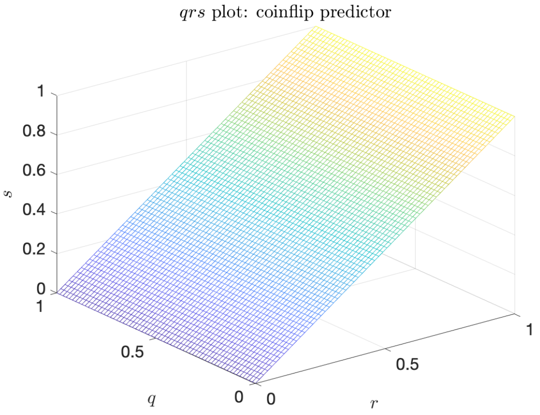


Figure 5. Plot of the qrs surface for the coinflip predictor. $V = 0.5$, $\alpha^* = 0$.

4.2. Bimodal Indicator—Predictor

A repository with sample codes for the following subsections may be found on Github. Sample implementations of the algorithms discussed may be found at https://doi.org/10.5281/zenodo.3417076, or https://github.com/batsteve/ferocious-cucumber. To demonstrate the relationship between separation of data (probability bimodality) and α^*, we consider a bimodal pdf, given by:

$$f_{ab}(a, b; \gamma, \rho) = \frac{1}{\beta}[e^{-(a^2+b^2)\rho^2} + \gamma e^{-((a-1)^2+(b-1)^2)\rho^2}] \quad (18)$$

This function is the sum of two Gaussian modes: a quiescent mode centered at $(0, 0)$ and an extreme event mode centered at $(1, 1)$. The pdf is further controlled by two parameters: γ and ρ. The parameter γ controls the weight of the extreme mode relative to the quiescent mode, and ρ controls the spacing of the modes. Figure 6 shows the joint pdf and the corresponding qrs plots for representative parameter values. Note that the knuckle becomes more pronounced as ρ (separation) increases.

The plots in Figure 7 shows the volume under the curve, V, and the maximum adjusted area under the curve, α^*, for the bimodal scenario as a function of γ and ρ. We note that V is largest when the separation between extreme and quiescent events, ρ, is large but γ is close to 1 (equal distribution of probability mass between extreme and and quiescent events). On the other hand, α^* peaks even at small values of γ, a scenario which is more realistic for extreme events occurring in physical systems due to internal instabilities, and which cannot be captured with the standard approach relying on V. This clearly demonstrates the advantage of the introduced criterion for the selection of optimal predictors.

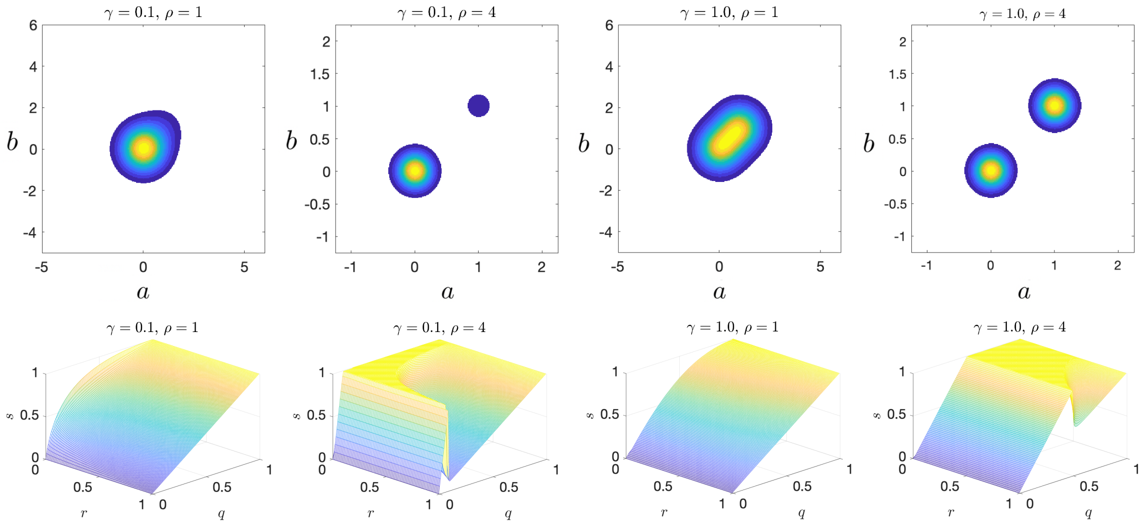


Figure 6. (**top**) Joint pdf plots of the bimodal scenario for various parameters; and (**bottom**) corresponding *qrs* plots.

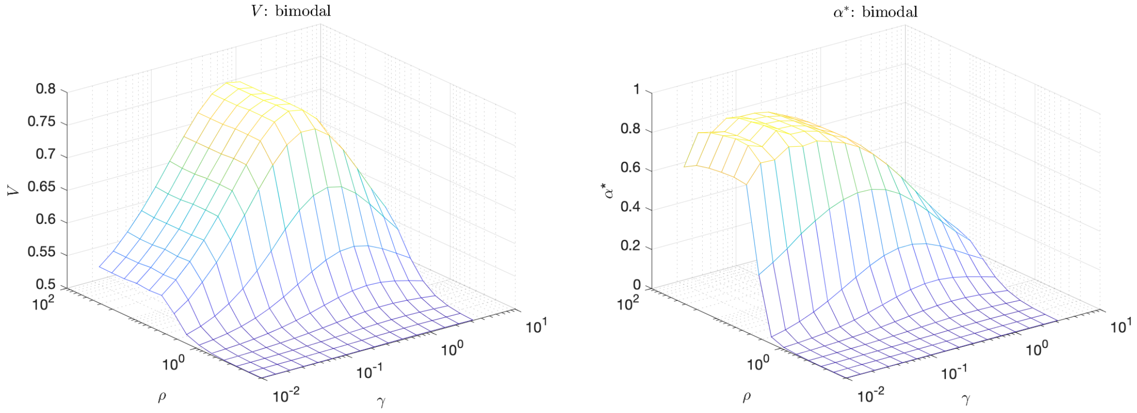


Figure 7. Volume under the curve (V) (**left**); and maximum adjusted area under the curve (α^*) (**right**) for the bimodal scenario.

4.3. Gaussian Indicator—Predictor

The multivariate Gaussian scenario is described by a pdf of the form

$$f_{ab}(a,b;\rho,\theta) = \frac{1}{\beta}e^{-(\cos\theta a+\sin\theta b)^2\frac{1}{\rho^2}-(\sin\theta a-\cos\theta b)^2\rho^2} \tag{19}$$

where β is a normalization factor, ρ^2 is the length ratio of the two principal axes, and θ is the angle between the principal axis and the a axis. Figure 8 shows a sample pdf. When $\theta = 0$ or $\theta = \frac{\pi}{2}$, there is no linear (or higher order) correlation between the indicator and predictor and in both cases the predictor is merely the coinflip predictor. However, when θ is near $\frac{\pi}{4}$, and $\rho > 1$, there *is* a (linear) correlation between a and b. In this regime, we see both V and α^* increase (Figure 9). Unlike the bimodal scenario, there is no distinct scale separation. As a result, V and α^* track each other closely.

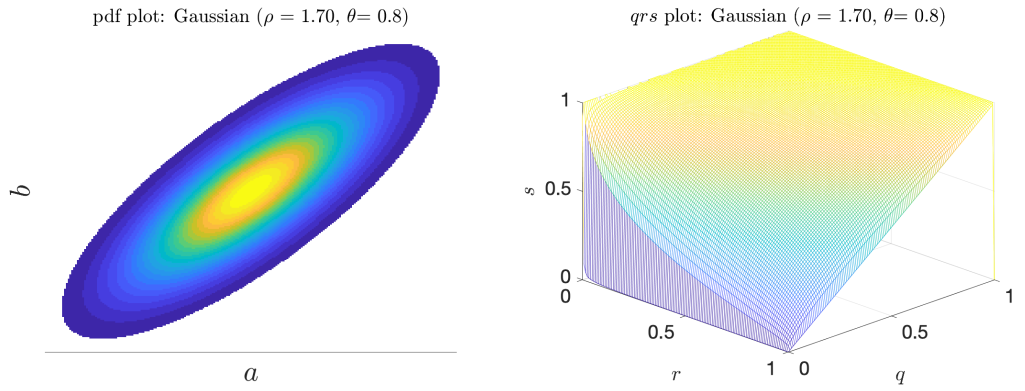


Figure 8. Joint pdf plot (**left**); and the corresponding qrs surface for the Gaussian scenario (**right**).

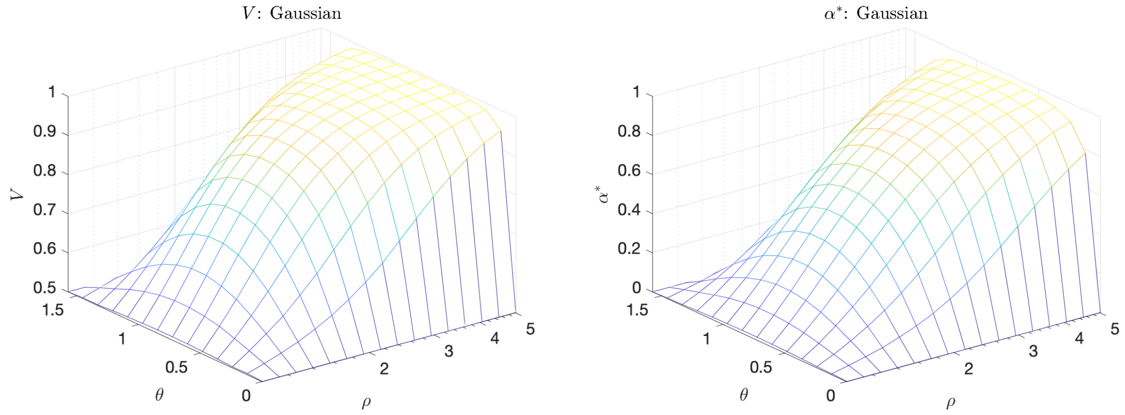


Figure 9. Volume Under the Curve (V) (**left**) and Maximum Adjusted Area Under the Curve (α^*) (**right**) for the Gaussian scenario.

5. Applications

5.1. The Majda–McLaughlin–Tabak (MMT) Model

The Majda–McLaughlin–Tabak (MMT) model is a 1D nonlinear model of deep water wave dispersion first introduced in [23], and since studied in the context of weak turbulence and intermittent extreme events [25,28,29]. The governing equation is given by

$$iu_t = |\partial_x|^\alpha u + \lambda |\partial_x|^{\frac{-\beta}{4}} \left(\left| |\partial_x|^{\frac{-\beta}{4}} u \right|^2 |\partial_x|^{\frac{-\beta}{4}} u \right) + iDu, \tag{20}$$

where u is a complex scalar, while the pseudodifferential operator $|\partial_x|^\alpha$ is defined through the Fourier transform as follows:

$$\widehat{|\partial_x|^\alpha u}(k) = |k|^\alpha \widehat{u}(k).$$

Here, we select the system parameters following the authors of [25], who employed the MMT equation to validate a model for nonlinear wave collapse and extreme event prediction. The domain has spatial extent 2π, discretized into 8192 points, and temporal extent 150, discretized into 6000 points for integration purposes. The parameters are chosen so that $\lambda = -4$ (focusing case), $\alpha = \frac{1}{2}$ (deep water waves case), and $\beta = 0$. The operator Du is a selective Laplacian designed to model dissipation at small scales, i.e., due to wave breaking. The first 1000 points are discarded to avoid transients due to random initial conditions; no forcing term is included and the simulations represent free decay.

Extreme events are indicated by large values of the wave group amplitude:

$$A(x_0, t_0) = |u(x_0, t_0)|. \tag{21}$$

Figure 10 shows sample realization of the MMT model near an extreme event.

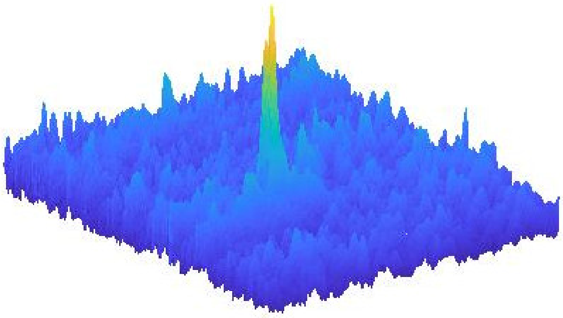
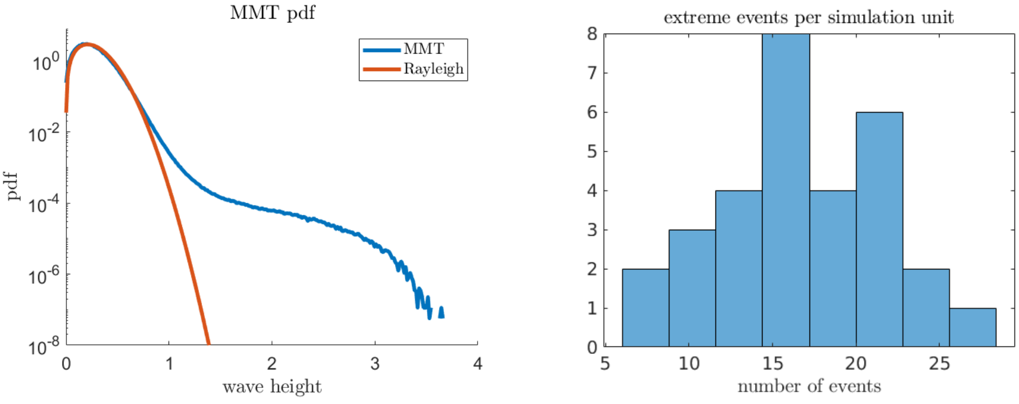


Figure 10. Sample plot of one simulated realization of the Majda–McLaughlin–Tabak (MMT) model near an extreme event.

We construct predictors using a simple machine learning paradigm: surrogate optimization across a specified hypothesis space. We compare α^* (introduced in Equation (17)) against other standard binary classification objectives, such as F_1-score and total accuracy. For total accuracy and F_1-score, we use an extreme-event threshold $\hat{a} = 1.5$ (Figure 11). The hypothesis class for the predictor is chosen as a two-element linear combination of k zero-order Gabor coefficients of variable length scales [30]:

$$h_P(x) = a_1 \exp\left(-\frac{x^2}{2L_1^2}\right) + a_2 \exp\left(-\frac{x^2}{2L_2^2}\right). \tag{22}$$

These functions can be conceptualized as localized Gaussian wavelets. We use a standard implementation of a surrogate search [31,32] to identify the optimal values for $\frac{a_1}{a_2}, L_1, L_2$. Except where otherwise noted, we terminated optimization after 100 function evaluations when calculating from 1 unit of simulation data.



(**a**) (**b**)

Figure 11. (**a**) Probability density function of the MMT wave height. Rayleigh distribution overlaid for comparison (note the "long-tail' extending from $x \approx 1.5$ to $x \approx 3.5$); and (**b**) histogram of the number of extreme events for different simulation runs.

In the MMT model, extreme events are localized both spatially and temporally. To convert from simulation data into training pairs, we use strict time-lags and spatial maxima. The rule strict time-lag means that a prediction from time t_0 is always paired to an indicator measured at exactly time $t_0 + \tau$. This underestimates predictor quality by throwing out "near-misses," but avoids the definitional quagmire needed to avoid the issue. The spatial maximization rule means that the most extreme prediction anywhere (at fixed time t_0) is compared to the most extreme event anywhere (at fixed time

$t_0 + \tau$). This overestimates goodness by including "false coincidences", however, in simulation, false coincidences are extremely rare.

Numerical Results

Regardless of the objective function, all the computed optimal predictors are typically characterized by a short length scale component, a long length scale component, and an amplitude weighting greatly favoring the short component. This breakdown has a simple physical interpretation that agrees with with previous work [25]. In order for an extreme event to occur, there must be sufficient background energy to draw up (long length scale), and also enough localized "seed" energy which will begin the collapse.

Exploratory investigations of three-component predictors (with five adjustable parameters) almost invariably collapsed onto two-vector solutions. This suggests that the two length-scale interpretation of the optimal predictors is not just a necessary artifact of hypothesis space dimension. The joint prediction–truth pdf (Figure 12a) does not appear to contain any scale separation. Indeed, the only easily visible feature is the density peak in the low-prediction low-indicator corner (true negatives). However, in the associated *qrs* surface plot (Figure 12b), the knuckle feature near $r = 0.05$ suggests that there is some hidden scale separation.

The different choices of objective function lead to different optimal parameters, as shown in Figure 13. Three of the objectives, F_1, V, and α^*, result in similar parameter values, while the total accuracy criterion, E_T, is quite different—generally resulting in longer length scales for the predictor.

A natural question is: How do the predictions (of the different optimal predictors) differ, and which one is better? Figure 14 shows two sets of ROC curves for the different optimal predictors. The precision–recall curve (Figure 14a) shows that the total-accuracy-optimized predictor can achieve slightly better precision at very low recall tolerances, but otherwise performs more poorly. In typical extreme event prediction contexts, a high recall (avoid false negatives) is a more valuable property.

The near-50% precision rate (Figure 14a) reflects an approximate temporal symmetry of the MMT extreme event mechanism: focusing and defocusing energy distributions look very similar, especially within a restricted hypothesis space that discards phase information.

In the previous numerical experiments, a fixed time gap $\tau = 0.015$ has been used to represent a suitable prediction time scale: long enough for significant wave evolution, short enough that good predictions are better than blind chance. Figure 15 shows the optimal predictor parameters associated with the α^* objective function for other choices of τ. The dramatic drop near $\tau = 0.007$ obscures the fact that the predictor parameters are interdependent–the increase of the amplitude ratio partially counteracts the effects of the shorter length scales.

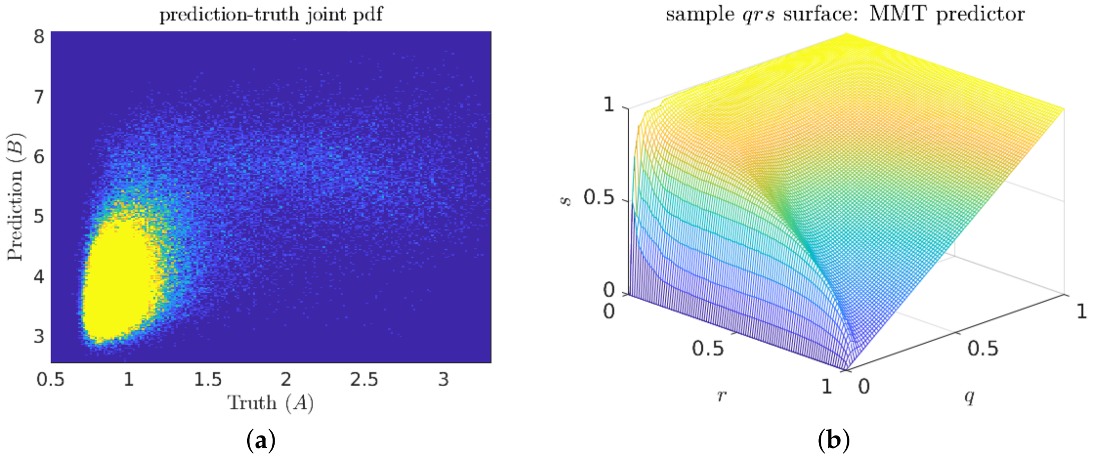


Figure 12. (**a**) Sample prediction–truth joint pdf for a good MMT predictor; and (**b**) corresponding *qrs* surface plot.

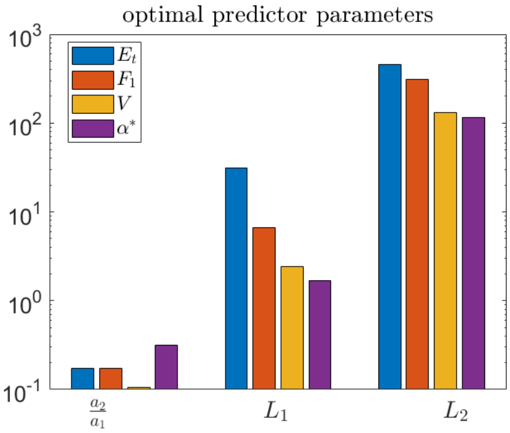


Figure 13. Optimal predictor parameters for each objective function. Note that total accuracy is very different from the others.

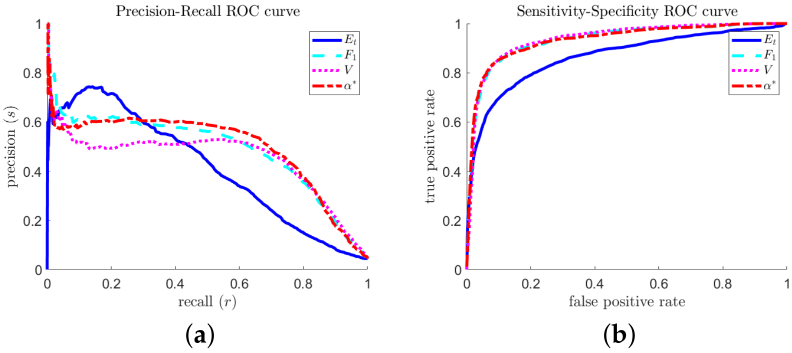


Figure 14. Receiver operating characteristic curve comparisons of optimal predictors calculated via different objectives: (**a**) precision–recall curve; and (**b**) sensitivity–specificity curve.

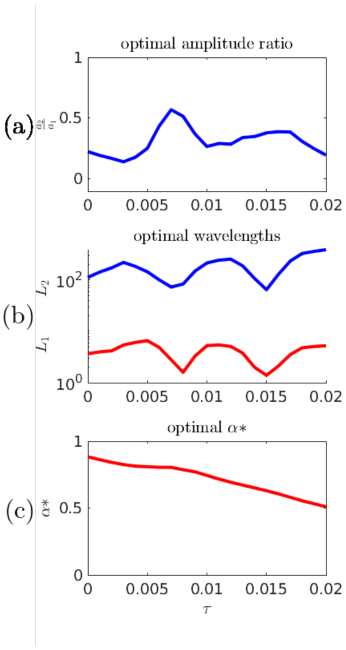


Figure 15. (**a**,**b**) Optimal predictor parameters as a function of τ; and (**c**) optimal α^* as a function of τ.

5.2. The Kolmogorov Flow Model

The Kolomogorov flow is a solution to the forced Navier–Stokes problem on a 2D periodic domain. Above Re $\approx$ 35, the solution is unstable, and there are intermittent bursts of energy dissipation [14]. The Navier–Stokes equations (pressure–velocity form), defined over some domain $\Omega \in \mathcal{R}^2$, are given by

$$\begin{aligned} \partial_t u &= -u \cdot \nabla u - \nabla p + \nu \Delta u + f \\ \nabla \cdot u &= 0 \end{aligned} \tag{23}$$

where u is the (vector valued) fluid velocity field, p is the (scalar valued) pressure field, ν is the dimensionless viscosity (inversely related to the famous Reynolds number) and f is some forcing term. In the Kolmogorov flow model, the forcing is a monochromatic time invariant field given by

$$f(\mathbf{x}) = \sin(\mathbf{k}_y \cdot x)\hat{e}_1 \tag{24}$$

where $\mathbf{k}_y = (0, 4)$ is the wavenumber of the forcing field and $\hat{e}_1 = (1, 0)$ is a unit vector perpendicular to $\mathbf{k}_y$. The intermittent bursting phenomena associated with the Kolmogorov flow for large enough Reynolds numbers (Re $\gtrapprox$ 35) are captured by the energy dissipation rate, given by

$$E_{\text{diss}}(u) = \frac{\nu}{|\Omega|} \int_\Omega |\nabla u|^2 dx \tag{25}$$

Figure 16 contains descriptive plots for the Kolmogorov flow model: two visualizations of a sample realization, and the pdf function.

A natural global set of predictors are the coefficients associated with low-k 2D Fourier modes, $b_{\mathbf{k}}$. We also consider arbitrary linear combinations of these coefficients, that is, predictors given by

$$B = \Sigma_{\mathbf{k}} \gamma_{\mathbf{k}} b_{\mathbf{k}} \tag{26}$$

Other machine learning algorithmic choices closely track those employed for the MMT model, although intermittent events are spatially global so there is no spatial maximization required.

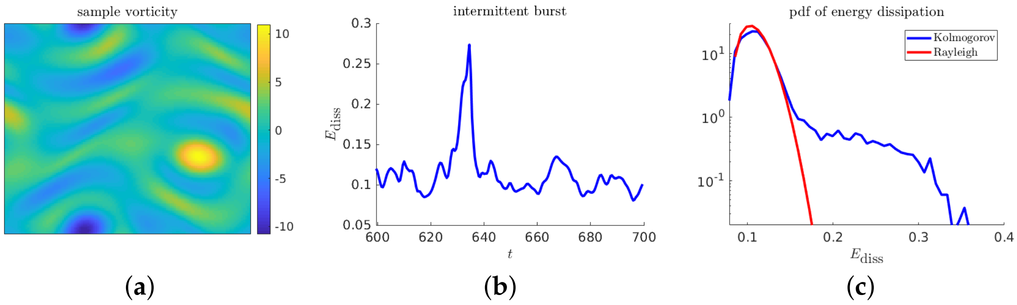


Figure 16. Descriptive plots for the Kolmogorov flow: (**a**) sample realization of the the vorticity; (**b**) time series of energy dissipation near an extreme event; and (**c**) pdf of the energy dissipation.

Numerical Results

Figure 17 shows the prediction quality of the single-coefficient predictors according to different metrics. By every metric, the Fourier coefficient with wavenumber $k = (0, 4)$ has consistently good predictive properties. This is consistent with the findings in [14] where a variational approach was employed to compute precursors directly from the governing equations and some basic properties of the system attractor. Figure 18 plots the six most significant $\gamma_{\mathbf{k}}$ from Equation (26) to show the effects of changing τ on the composition of the optimal predictors.

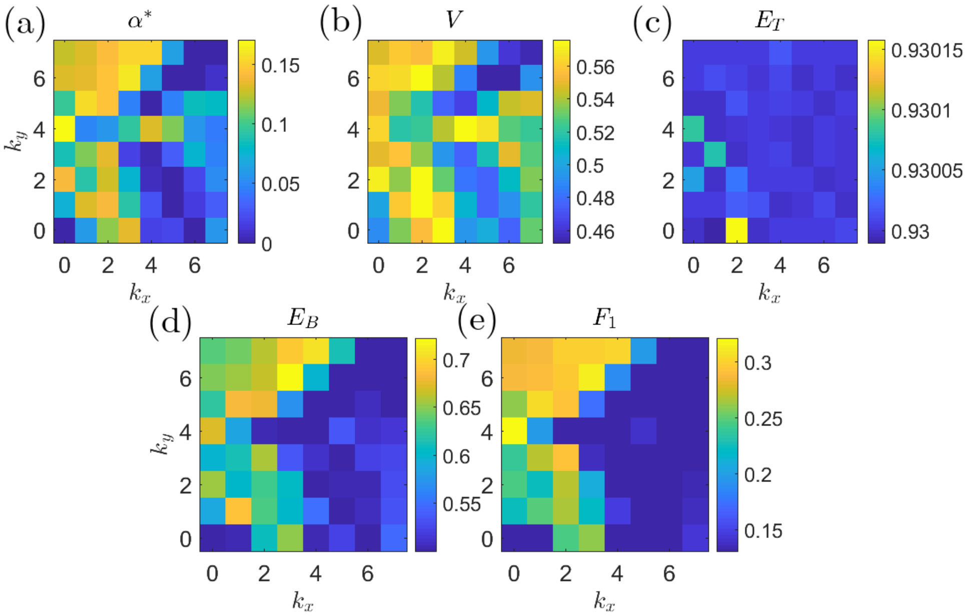


Figure 17. Plots of single coefficient predictor quality for different wavenumbers and objectives: (**a**) α^*; (**b**) volume under the surface; (**c**) total accuracy; (**d**) balanced error; and (**e**) F_1 score. Note the consistent peak at $(0,4)$, which is resolved best by α^* and F_1.

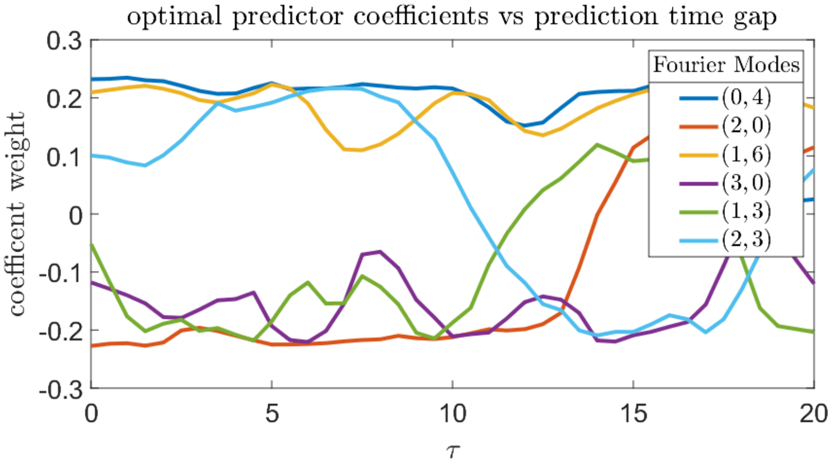


Figure 18. Composition of optimal predictor, in terms of Fourier modes as a function of prediction gap τ. Positive coefficients correspond to increasing extreme event likelihood, while negative coefficients correspond to suppression.

The next numerical study involves optimization of a combined predictor employing several wavenumbers. Even in this case, the most important component from the combined predictor is the $(0,4)$ mode, and its weighting factor is always positive. Other Fourier modes, such as $(3,0)$, have a negative weighting factor, which means they are inversely correlated with bursts of extreme dissipation. Due to the consistent downward trend in prediction quality as τ increases, trends in the data past $\tau \approx 15$ are less likely to be meaningful.

6. Conclusions

We have formulated a method for optimizing extreme event prediction in an equation free manner, i.e., using only data. Our first aim was to understand critical limitations of binary classification methods in the context of precursors for extreme events. We then showed how the *qrs* surface construction allows for a geometric interpretation of scale separation, and naturally led to the metric α^*, which is well suited to this problem. We compared α^* to other metrics in two models of extreme events, where we showed that α^* selects for qualitatively better predictions than the total accuracy, and has superior optimization properties as compared to F_1-score.

While the main focus of this work was a way to construct a prediction metric suited to the problem of extreme event prediction, there are still important questions related to the selection of good hypothesis classes, and the binning procedure to go from trajectory data to training pairs. While we believe both these questions are highly problem dependent, any attempt to apply the machine learning paradigm to a related problem must focus on these issues carefully.

Author Contributions: S.G. and T.P.S. conceptualized the work. S.G. developed the theory, carried out the simulations, and performed the computations. T.P.S. supervised the findings of the work, including suggestions for formal analysis. S.G. and T.P.S. analyzed the results and contributed to the final manuscript.

Funding: This work was funded by the Army Research Office (Grant No. W911NF-17-1-0306) and the Office of Naval Research (Grant No. N00014-15-1-2381).

Acknowledgments: The authors would like to thank the Editors for the invitation to this Special Issue. TPS also acknowledges support from a Doherty Career Development Chair at MIT.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chaloner, K.; Verdinelli, I. Bayesian Experimental Design: A Review. *Stat. Sci.* **1995**, *10*, 273–304. [CrossRef]
2. Higham, D.J. An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations. *SIAM Rev.* **2001**, *43*, 525–546. [CrossRef]
3. Dysthe, K.; Krogstad, H.E.; Müller, P.; Muller, P. Oceanic Rogue Waves. *Annu. Rev. Fluid Mech.* **2008**, *40*, 287. [CrossRef]
4. Kharif, C.; Pelinovsky, E.; Slunyaev, A. Rogue Waves in the Ocean, Observation, Theories and Modeling. In *Advances in Geophysical and Environmental Mechanics and Mathematics Series*; Springer: Berlin, Germany, 2009; Volume 14.
5. Li, F. Modelling the Stock Market Using a Multi-Scale Approach. Master's Thesis, University of Leicester, Leicester, UK, 2017.
6. Akiko Kashiwagi, Itaru Urabe, K.K.; Yomo, T. Adaptive Response of a Gene Network to Environmental Changes by Fitness-Induced Attractor Selection. *PLoS ONE* **2006**, *1*, 1–10. [CrossRef] [PubMed]
7. Zio, E.; Pedroni, N. Estimation of the functional failure probability of a thermal-hydraulic passive system by Subset Simulation. *Nucl. Eng. Des.* **2009**, *239*, 580–599. [CrossRef]
8. Beibei X.; Feifei W.; Diyi, C.; Zhang, H. Hamiltonian modeling of multi-hydro-turbine governing systems with sharing common penstock and dynamic analyses under shock load. *Energy Convers. Manag.* **2016**, *108*, 478–487.
9. Varadhan, S.R.S. *Large Deviations and Applications*; SIAM: Philadelphia, PA, USA, 1984.
10. E, W.; Vanden-Eijnden, E. Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.* **2010**, *61*, 391–420. [CrossRef]
11. Qi, D.; Majda, A.J. Predicting Fat-Tailed Intermittent Probability Distributions in Passive Scalar Turbulence with Imperfect Models through Empirical Information Theory. *Commun. Math. Sci.* **2016**, *14*, 1687–1722. [CrossRef]
12. Mohamad, M.A.; Sapsis, T.P. Probabilistic Description of Extreme Events in Intermittently Unstable Dynamical Systems Excited by Correlated Stochastic Processes. *SIAM/ASA J. Uncertain. Quantif.* **2015**, *3*, 709–736. [CrossRef]
13. Majda, A.J.; Moore, M.N.J.; Qi, D. Statistical dynamical model to predict extreme events and anomalous features in shallow water waves with abrupt depth change. *Proc. Natl. Acad. Sci. USA* **2018**, *116*, 3982–3987. [CrossRef]
14. Farazmand, M.; Sapsis, T.P. A variational approach to probing extreme events in turbulent dynamical systems. *Sci. Adv.* **2017**, *3*, e1701533. [CrossRef] [PubMed]

15. Wan, Z.Y.; Vlachas, P.R.; Koumoutsakos, P.; Sapsis, T.P. Data-assisted reduced-order modeling of extreme events in complex dynamical systems. *PLoS ONE* **2018**. [CrossRef] [PubMed]
16. Mohamad, M.A.; Sapsis, T.P. Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 11138–11143. [CrossRef] [PubMed]
17. Viv Bewick, L.C.; Ball, J. Statistics review 13: Receiver operating characteristic curves. *Crit. Care* **2004**, *8*, 508–512. [CrossRef] [PubMed]
18. He, H.; Garcia, E. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *29*, 1263–1284.
19. Takaya Saito, M.R. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, 1–21.
20. Cousins, W.; Sapsis, T.P. Reduced-order precursors of rare events in unidirectional nonlinear water waves. *J. Fluid Mech.* **2016**, *790*. [CrossRef]
21. Farazmand, M.; Sapsis, T.P. Reduced-order prediction of rogue waves in two-dimensional deep-water waves. *J. Comput. Phys.* **2017**, *340*, 418–434. [CrossRef]
22. Dematteis, G.; Grafke, T.; Vanden-Eijnden, E. Rogue Waves and Large Deviations in Deep Sea. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 855–860. [CrossRef]
23. Majda, A.J.; McLaughlin, D.W.; Tabak, E.G. A one-dimensional model for dispersive wave turbulence. *J. Nonlinear Sci.* **1997**, *7*, 9–44. [CrossRef]
24. Platt, N.; Sirovich, L.; Fitzmaurice, N. An investigation of chaotic Kolmogorov flows. *Phys. Fluids A* **1991**, *3*, 681–696. [CrossRef]
25. Cousins, W.; Sapsis, T.P. Quantification and prediction of extreme events in a one-dimensional nonlinear dispersive wave model. *Physica D* **2014**, *280*, 48–58. [CrossRef]
26. Mohamad, M.; Sapsis, T. Probabilistic response and rare events in Mathieu's equation under correlated parametric excitation. *Ocean Eng.* **2016**, *120*. [CrossRef]
27. Mohamad, M.A.; Cousins, W.; Sapsis, T.P. A probabilistic decomposition-synthesis method for the quantification of rare events due to internal instabilities. *J. Comput. Phys.* **2016**, *322*, 288–308. [CrossRef]
28. David, C., Andrew, J.; McLaughlin, W.; Esteban, G.T. Dispersive wave turbulence in one dimension. *Physica D* **2001**, *152–153*, 551–572.
29. Benno Rumpf, L.B. Weak turbulence and collapses in the Majda–McLaughlin–Tabak equation: Fluxes in wavenumber and in amplitude space. *Physica D* **2005**, pp. 188–203. [CrossRef]
30. Gabor, D. Theory of Communication. *J. Inst. Electr. Eng.* **1946**, *93*, 429–457. [CrossRef]
31. Wang, S.; Metcalfe, G.; Stewart, R.L.; Wu, J.; Ohmura, N.; Feng, X.; Yang, C. Solid–liquid separation by particle-flow-instability. *Energy Environ. Sci.* **2014**, *7*, 3982–3988. [CrossRef]
32. Vu Khac Ky, Claudia D'Ambrosio, Y.H.; Liberti, L. Surrogate-based methods for black-box optimization. *Int. Trans. Oper. Res.* **2016**, *24*. [CrossRef]